Data science for lazy people, Automated Machine Learning

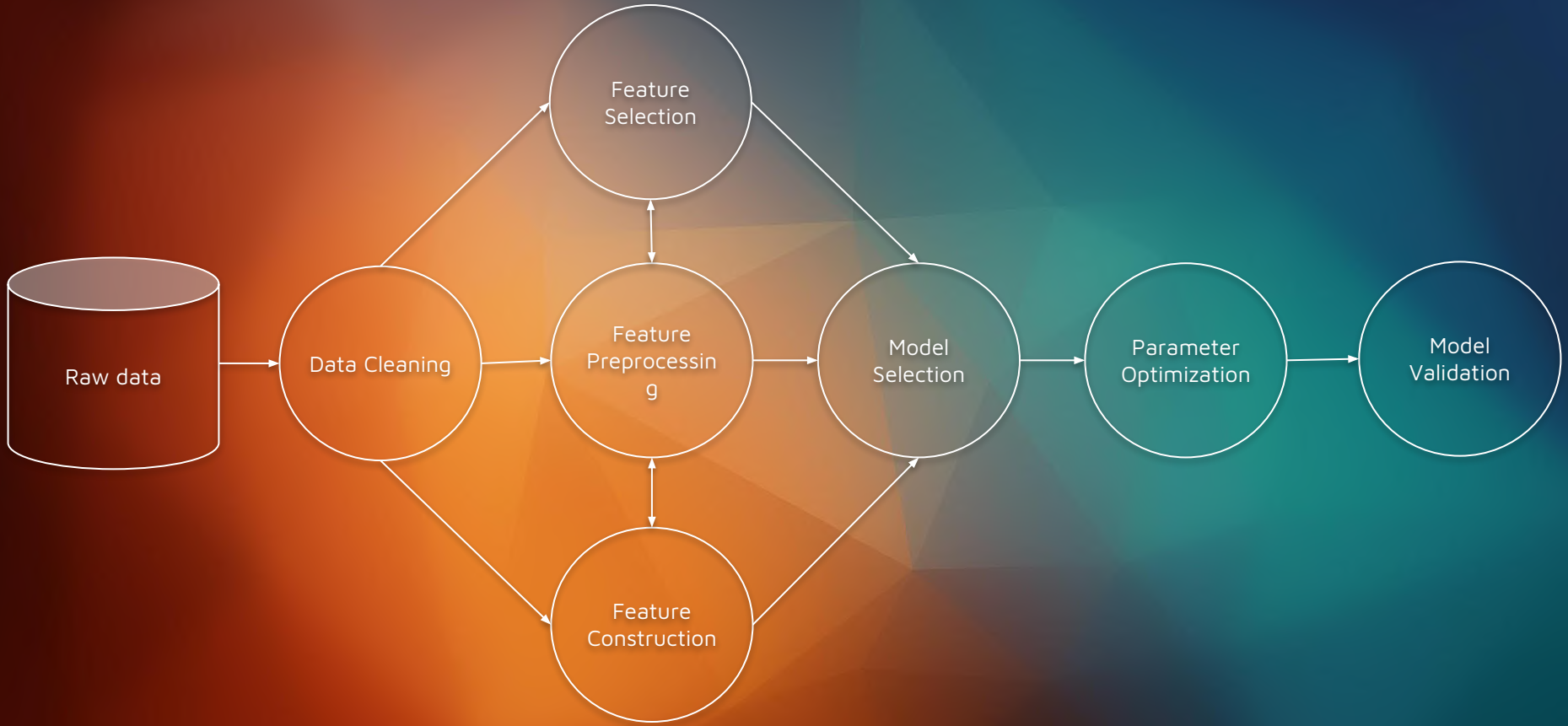
Big Data Congress Lithuania 2018

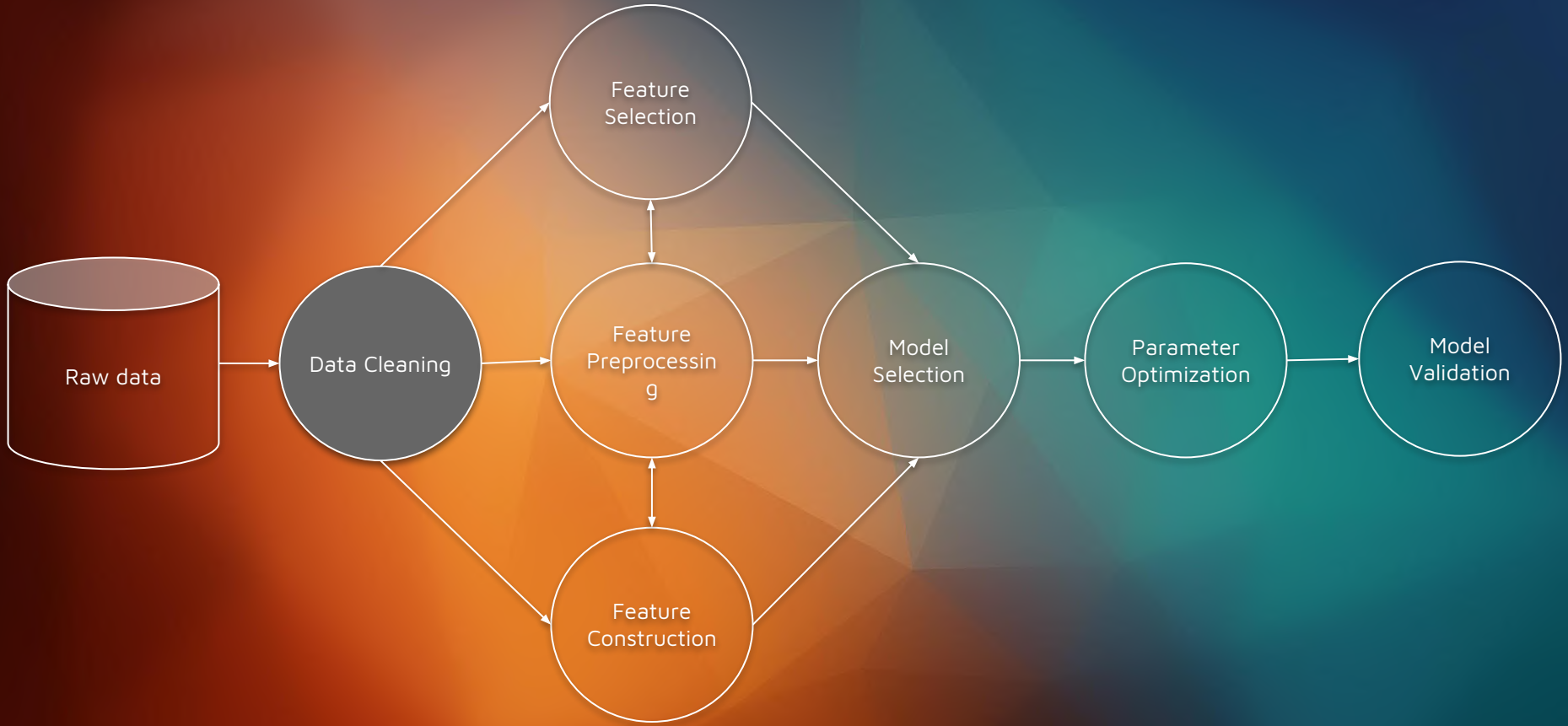
Diego Hueltes

 @jdiegoh

https://www.sli.do

#bigdata2018


Feature Selection
Raw data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction

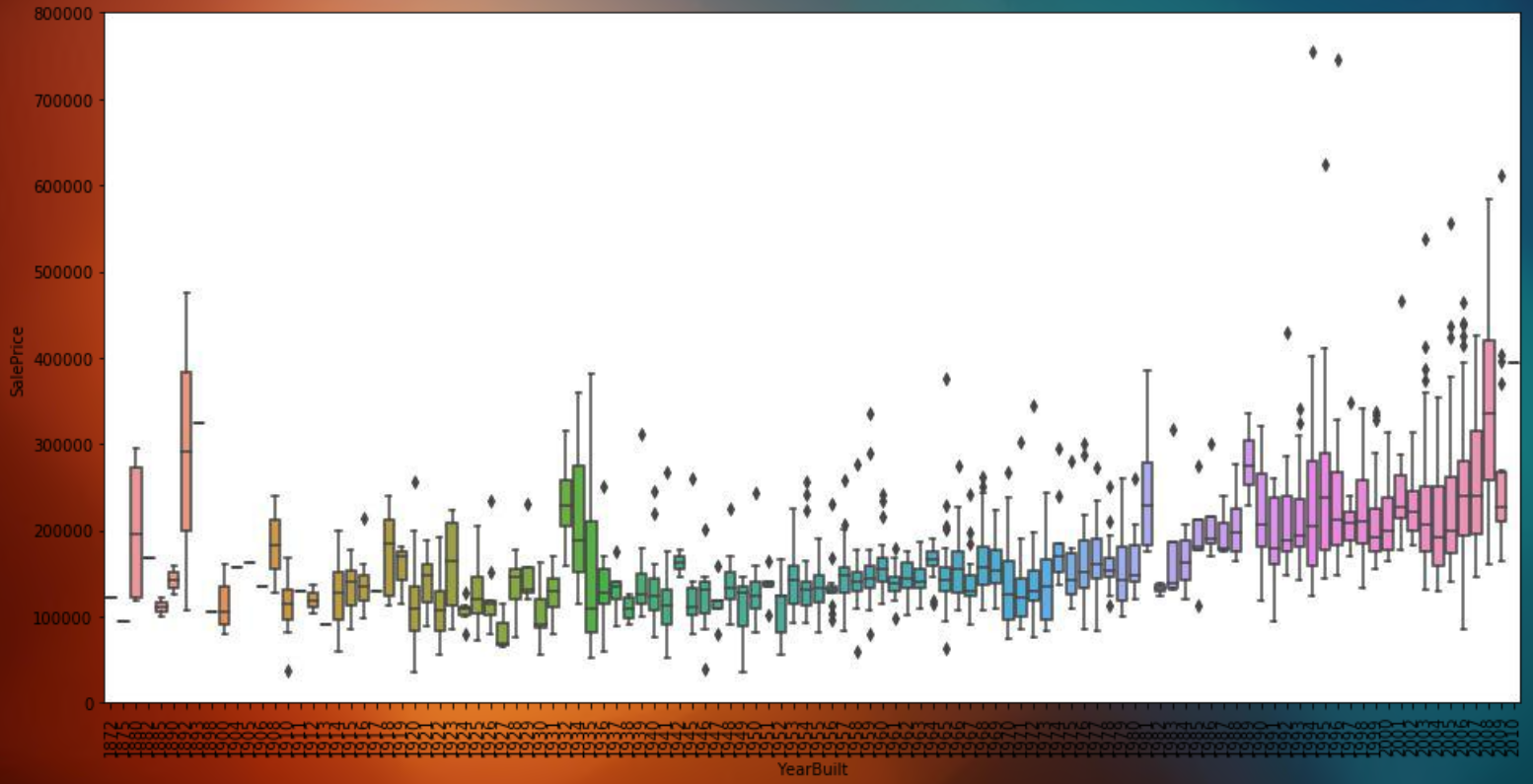

Raw data
Data Cleaning
Feature Selection
Feature Preprocessing
Feature Construction
Model Selection
Parameter Optimization
Model Validation

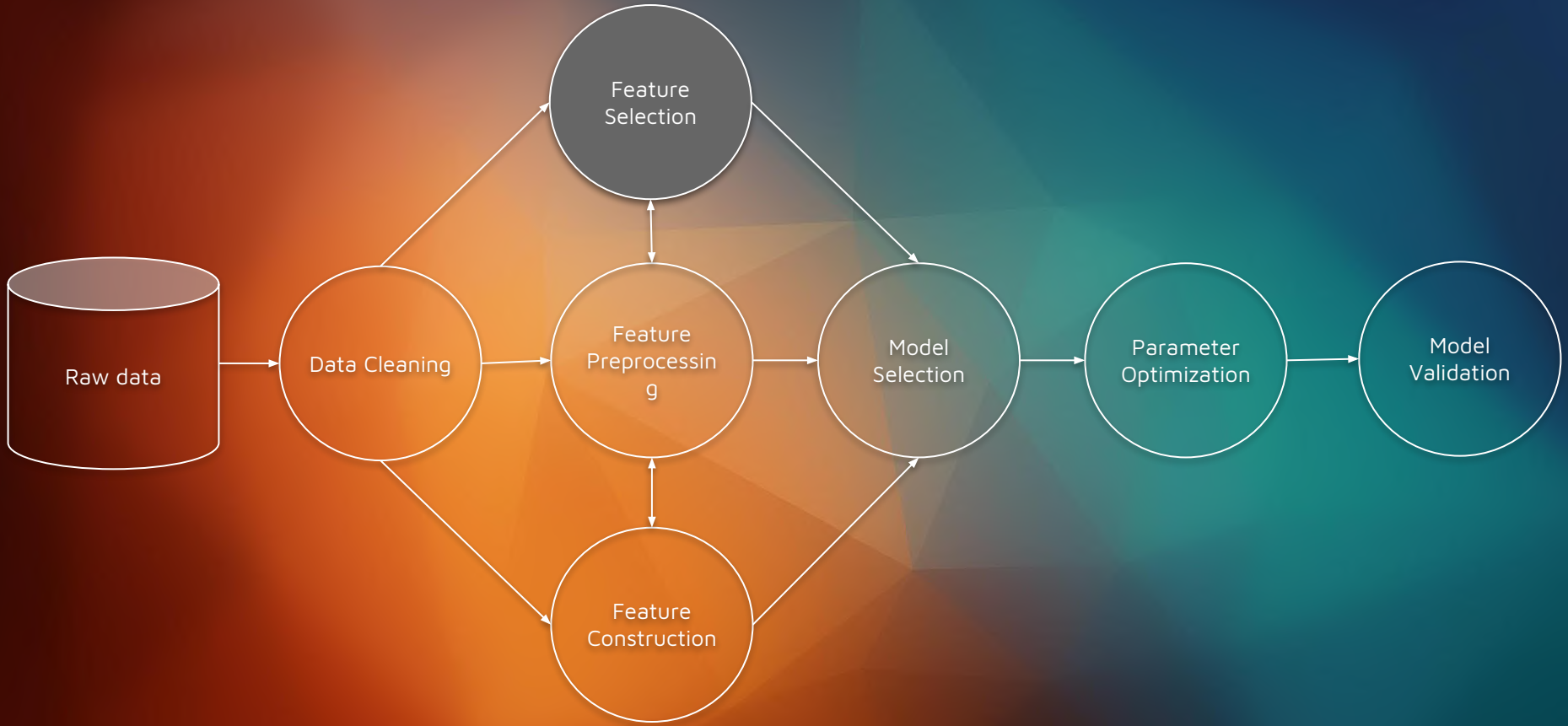
Data cleaning

Data cleaning

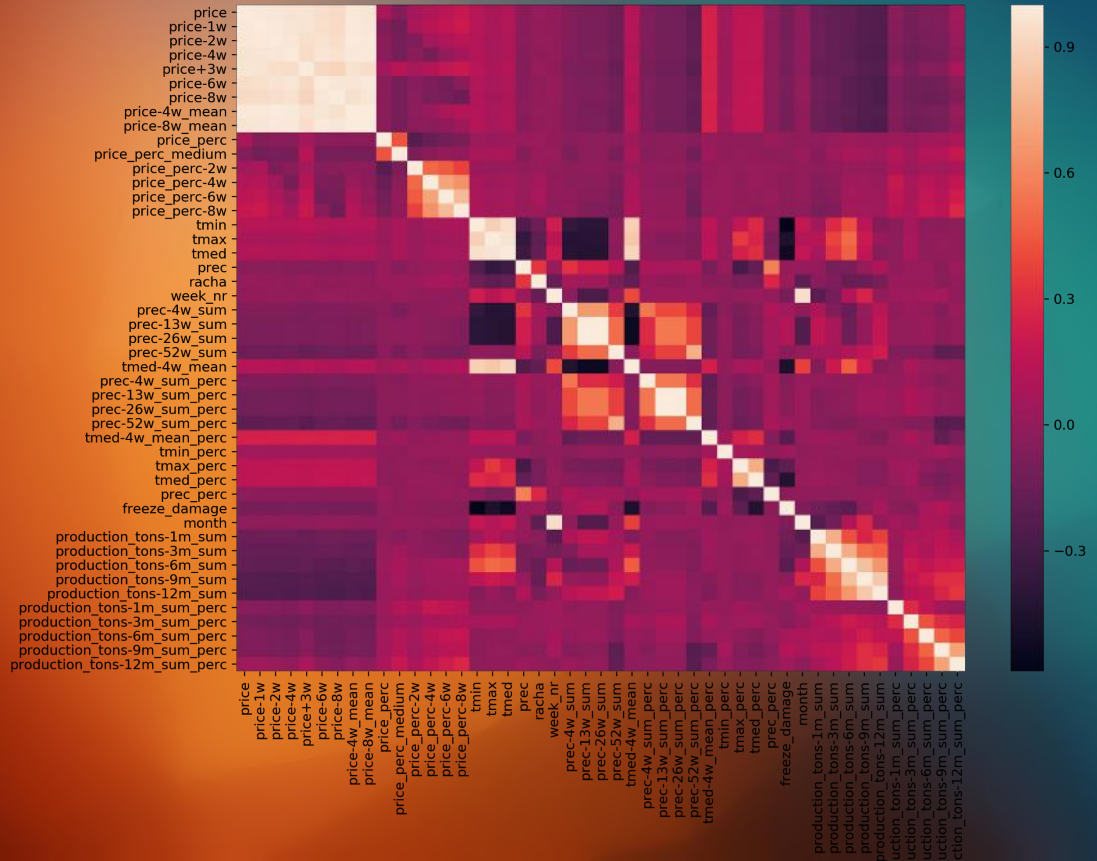
4.9	3.1	1.5	NA	Iris-setosa
5.4	3.7	1.5	NA	Iris-setosa
4.8	3.4	1.6	NA	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
5.7	3.0	1.1	0.1	Iris-setosa
5.8	4.0	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setossa
5.1	3.8	1.5	0.3	Iris-setosa

Data cleaning

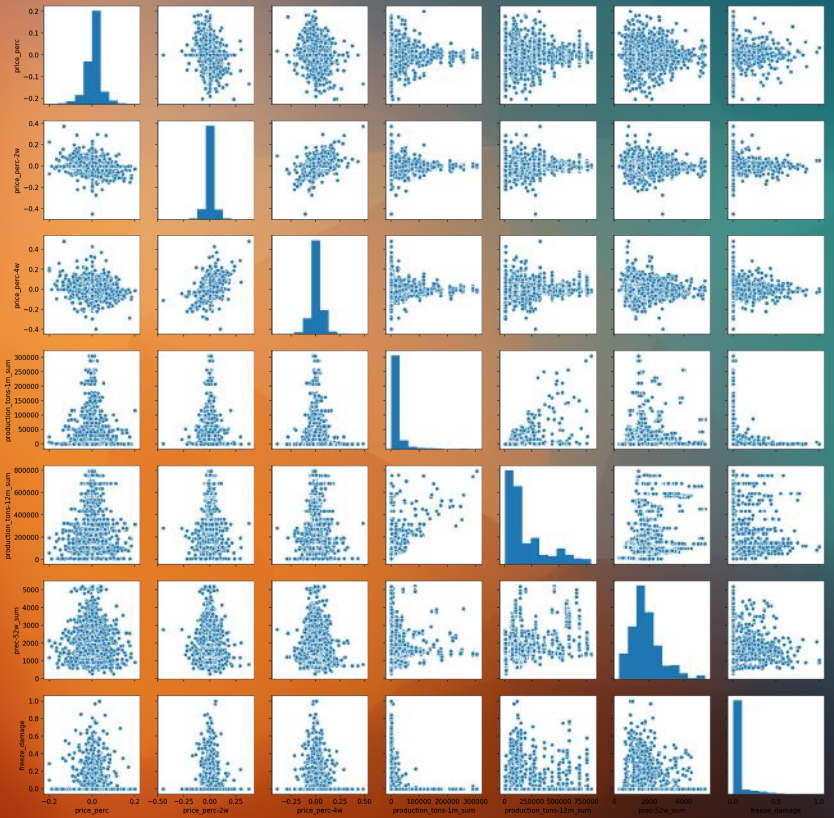


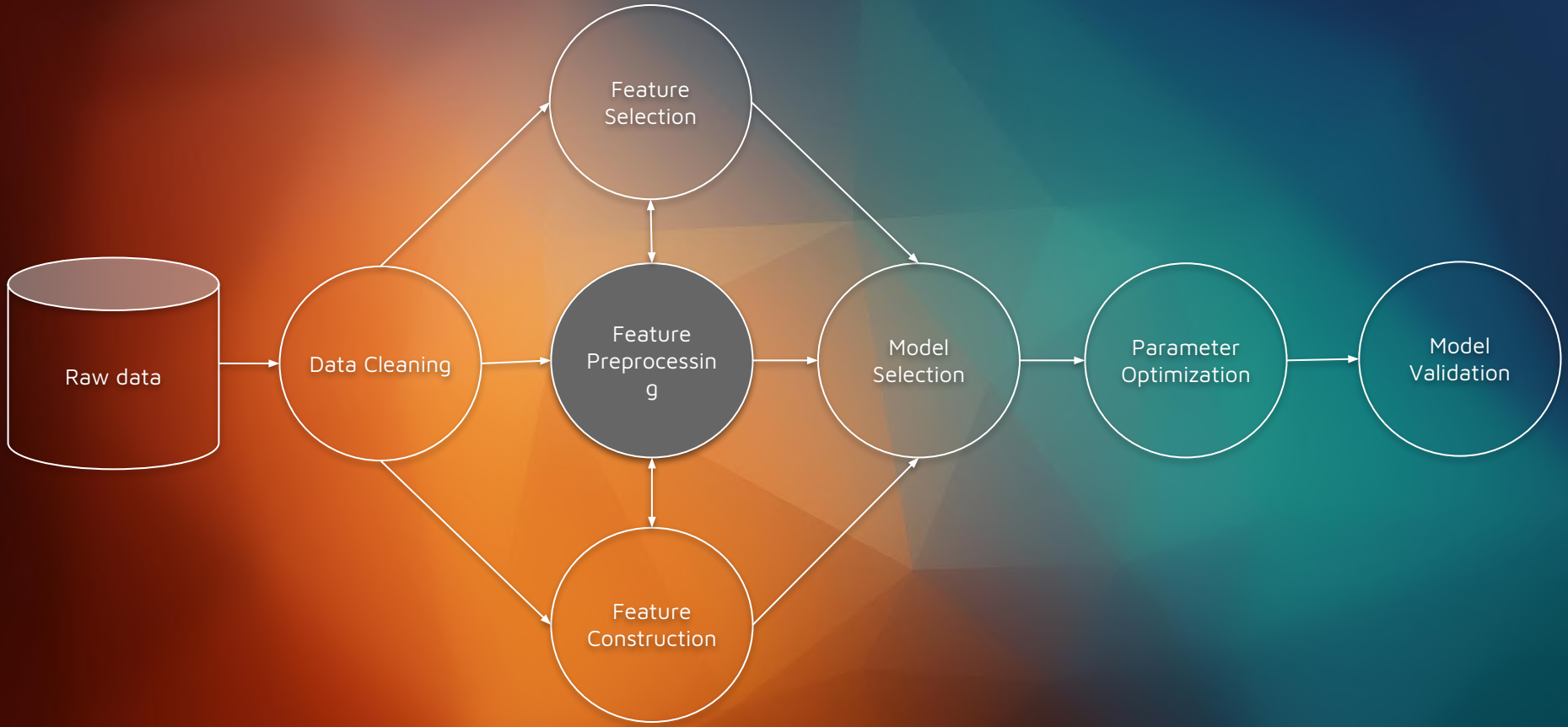

Feature Selection
Raw data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction

Feature selection

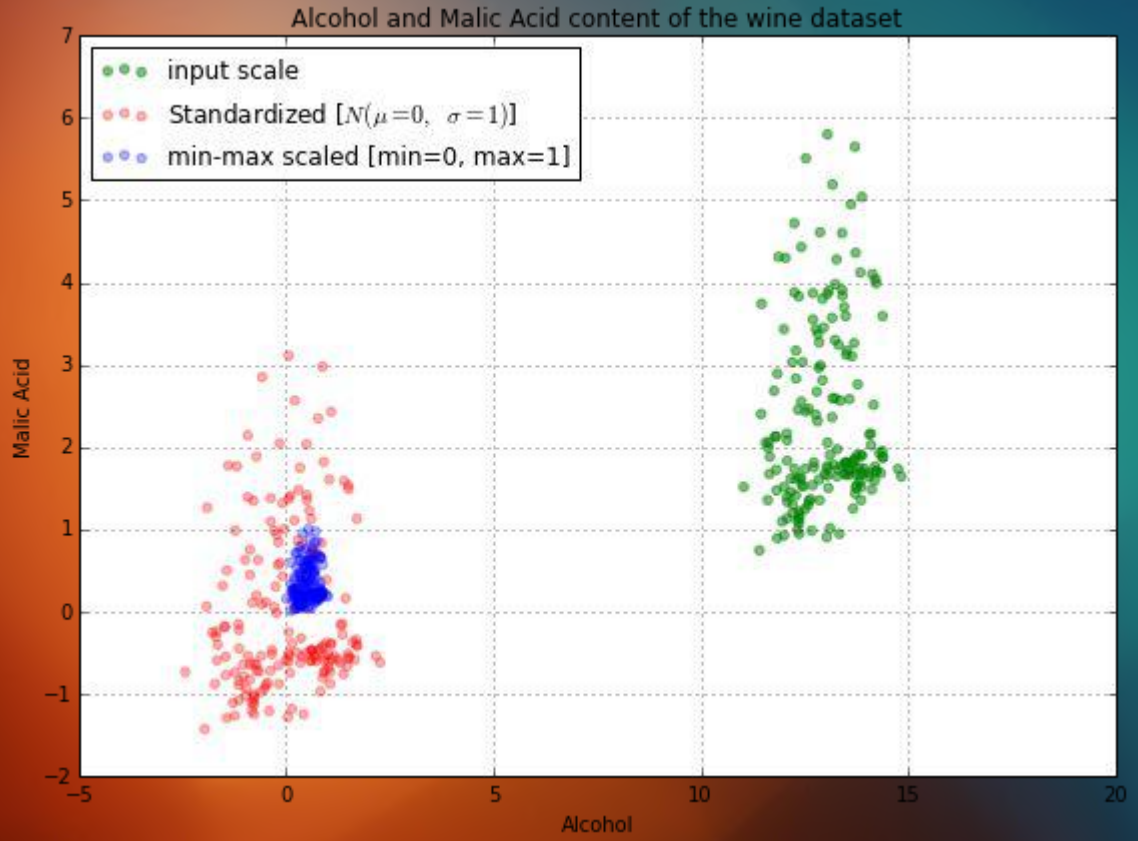


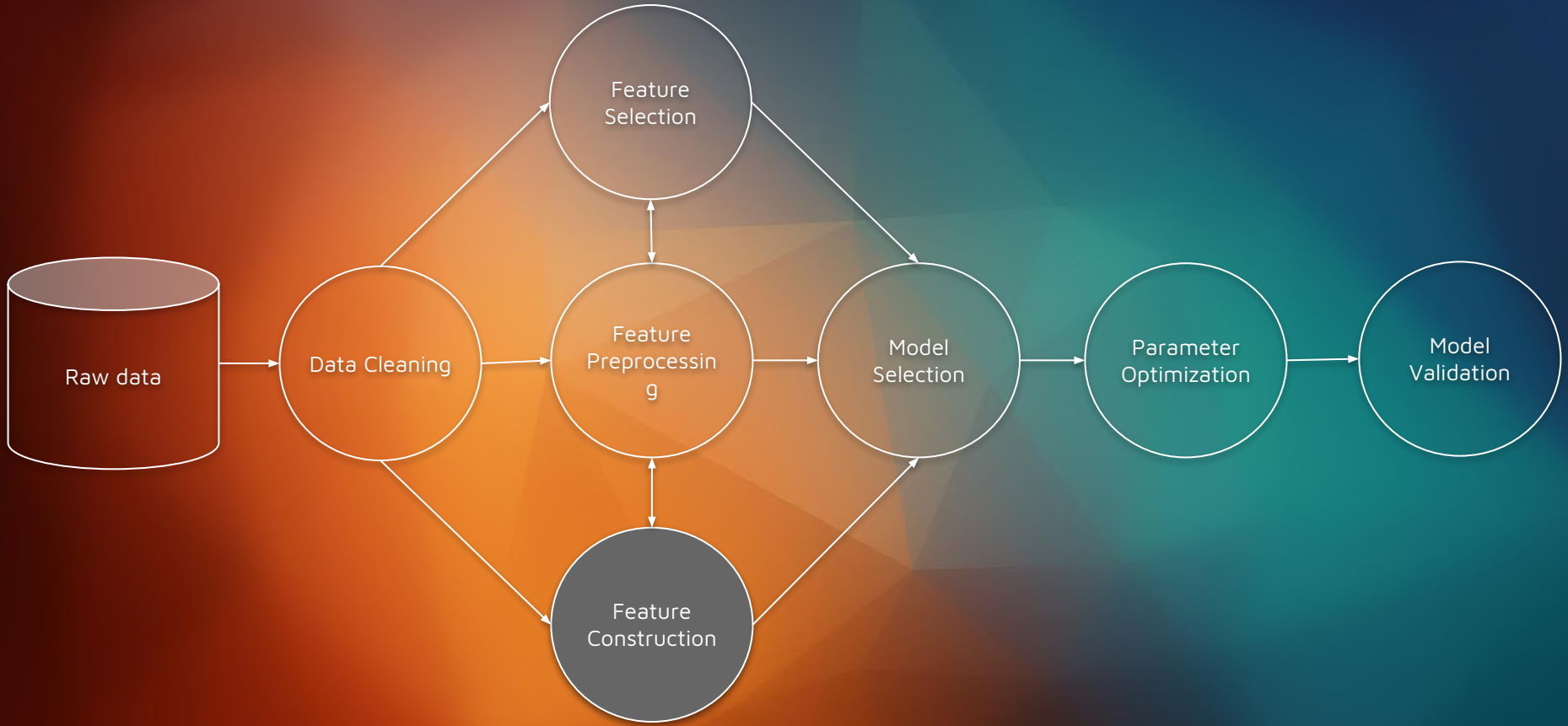
Feature selection



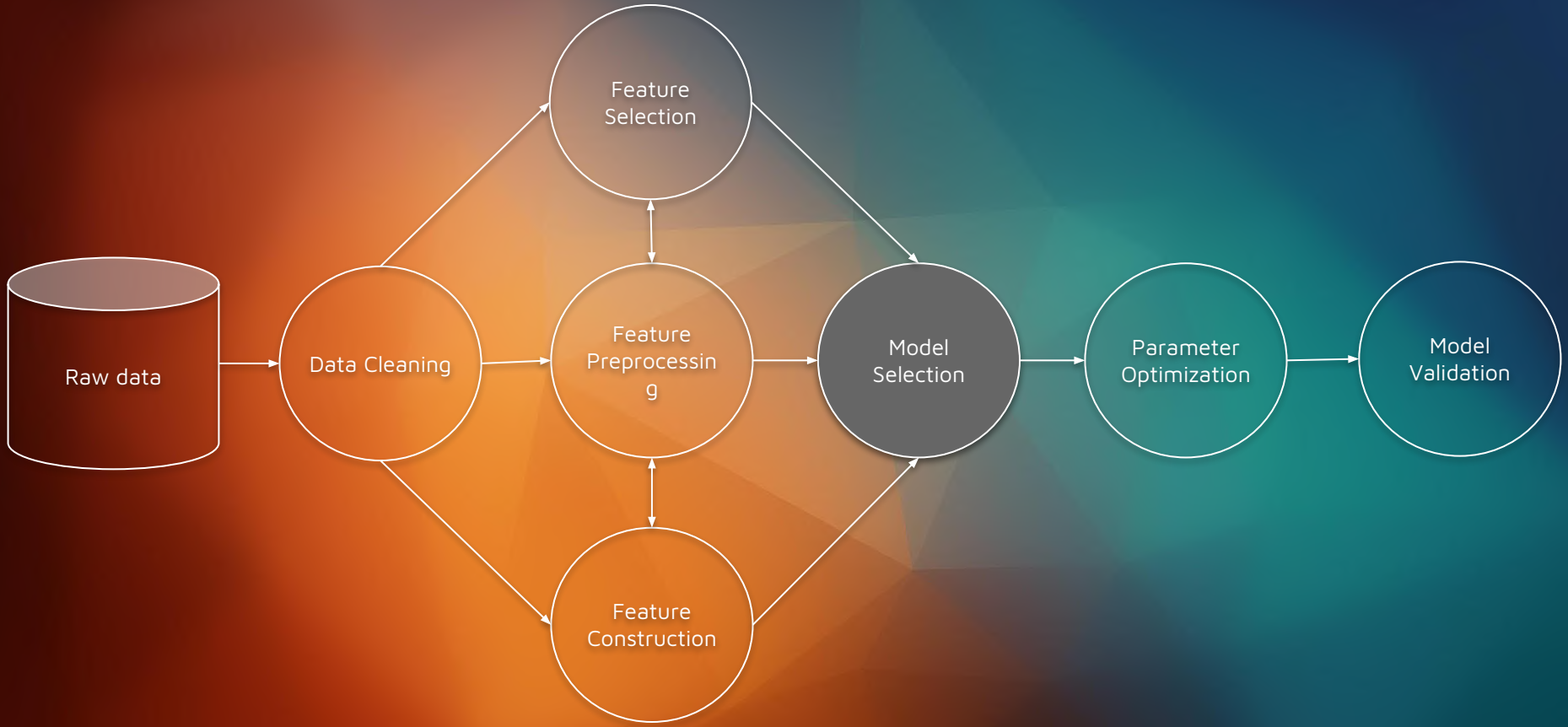

Raw data
Data Cleaning
Feature Selection
Feature Preprocessing
Feature Construction
Model Selection
Parameter Optimization
Model Validation

Feature preprocessing

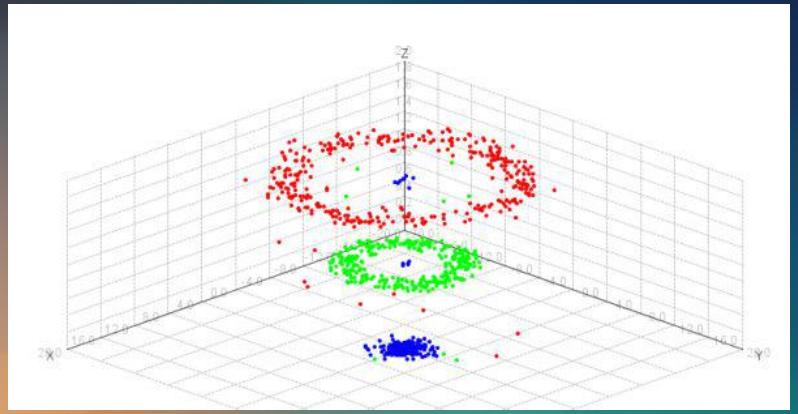


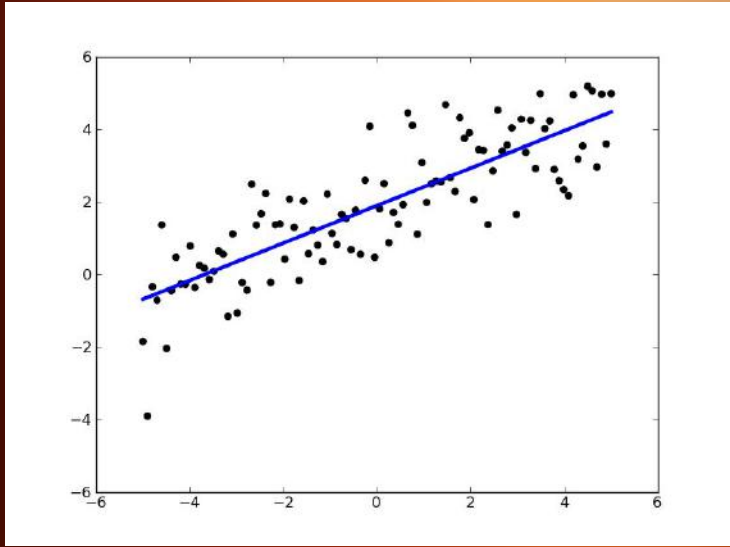

Feature Selection
Raw data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction

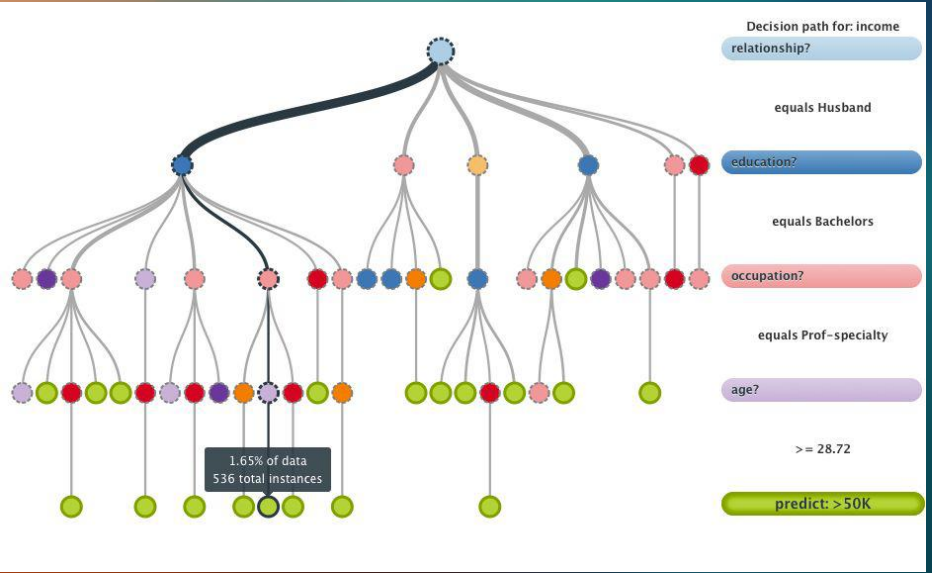
Feature construction

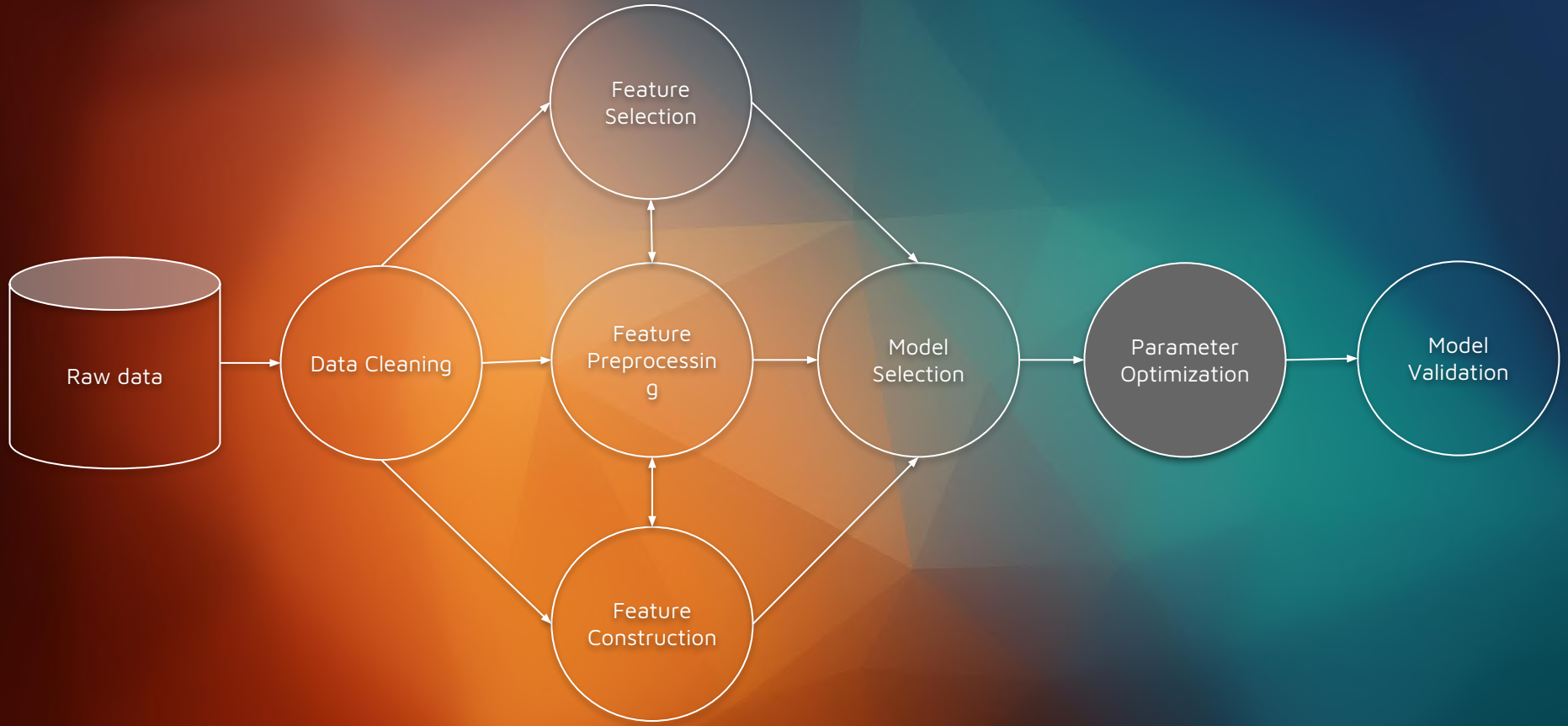

Feature Selection
Raw data
Data Cleaning
Feature Preprocessin g
Model Selection
Parameter Optimization
Model Validation
Feature Construction

Model selection



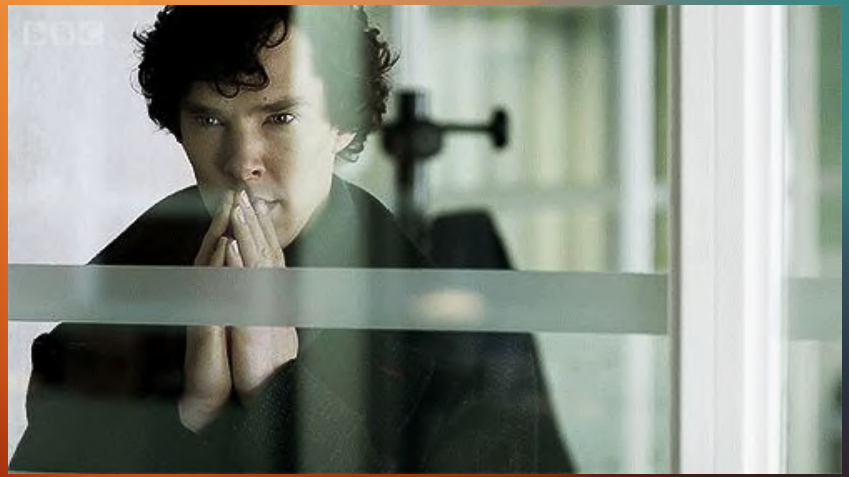


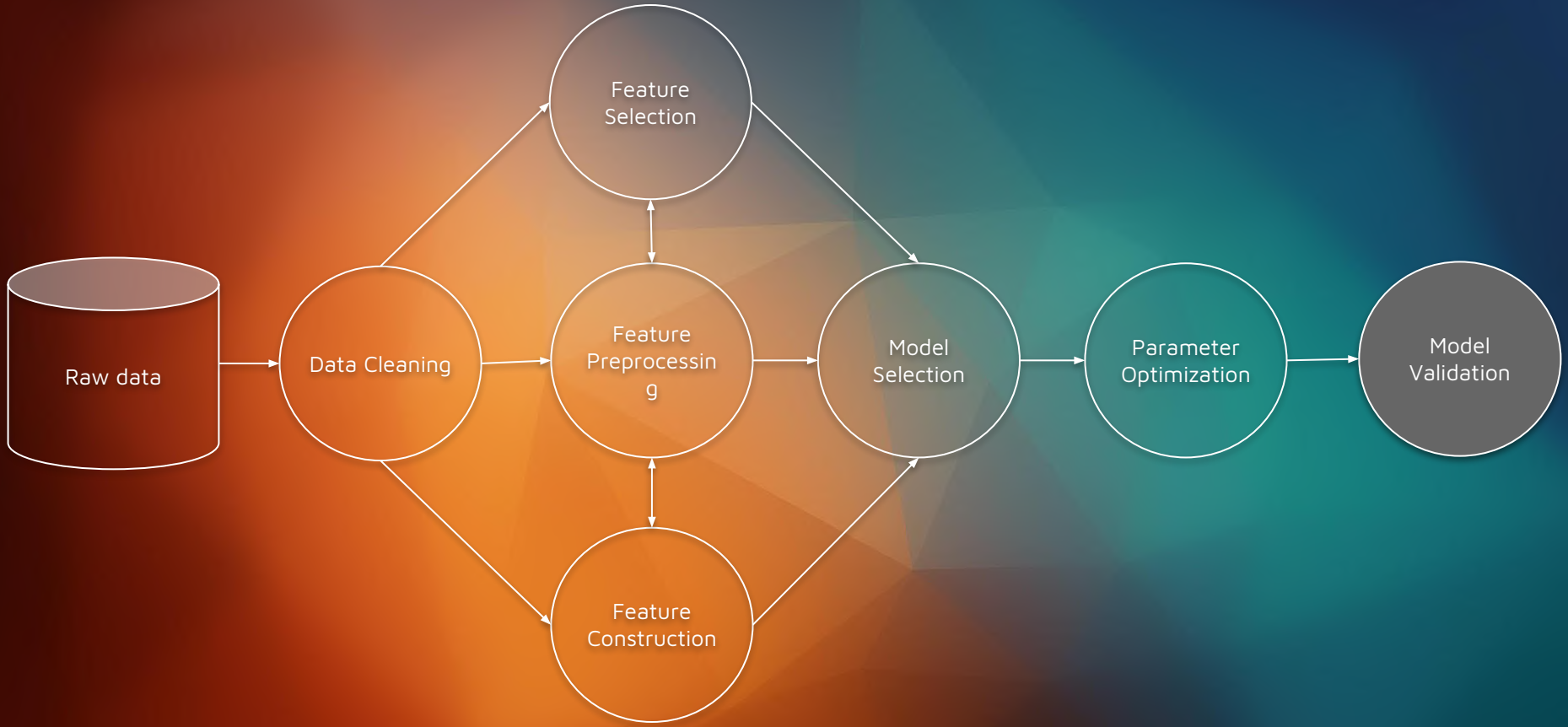



Feature Selection
Raw data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction

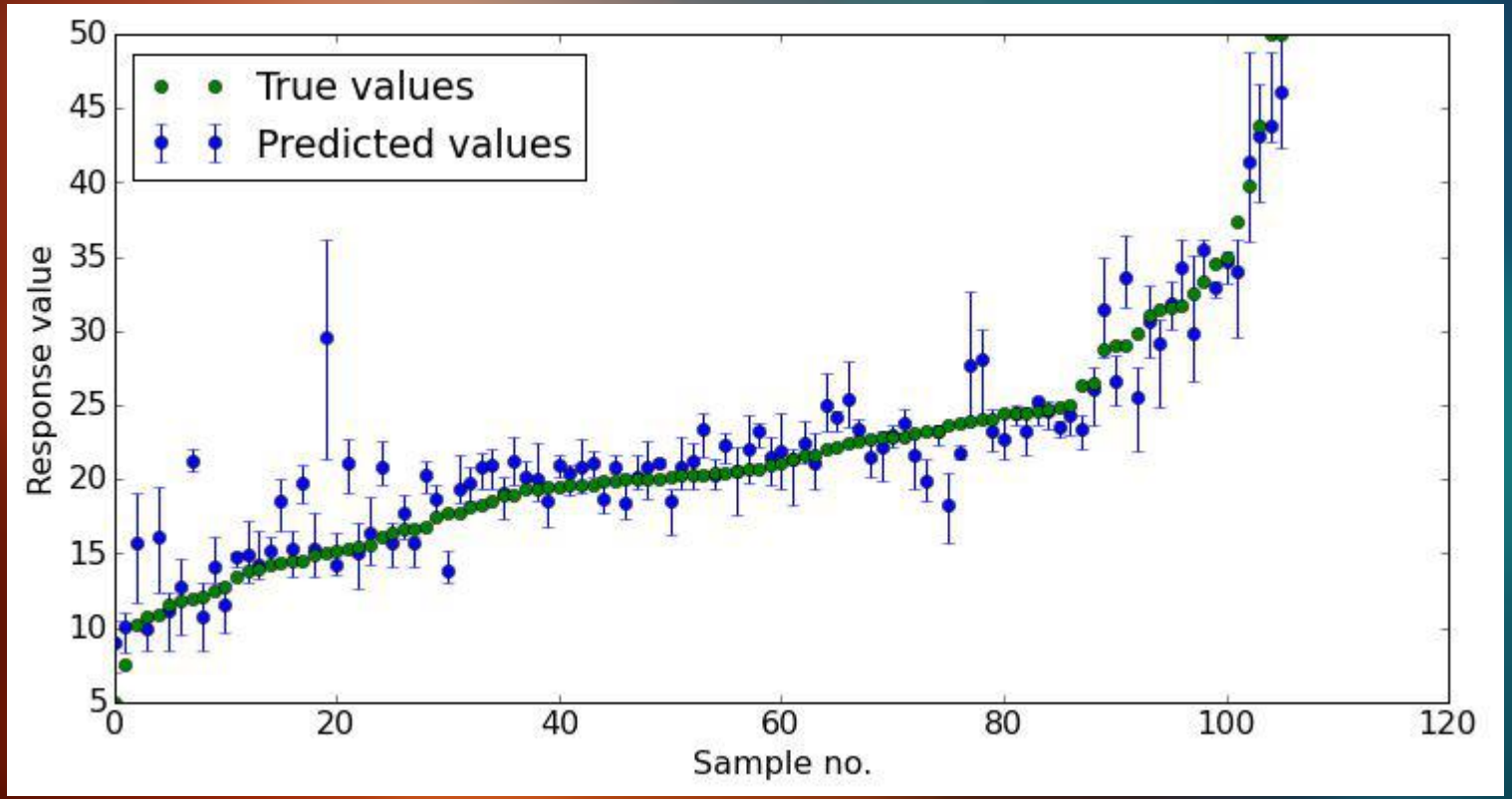
Parameter optimization

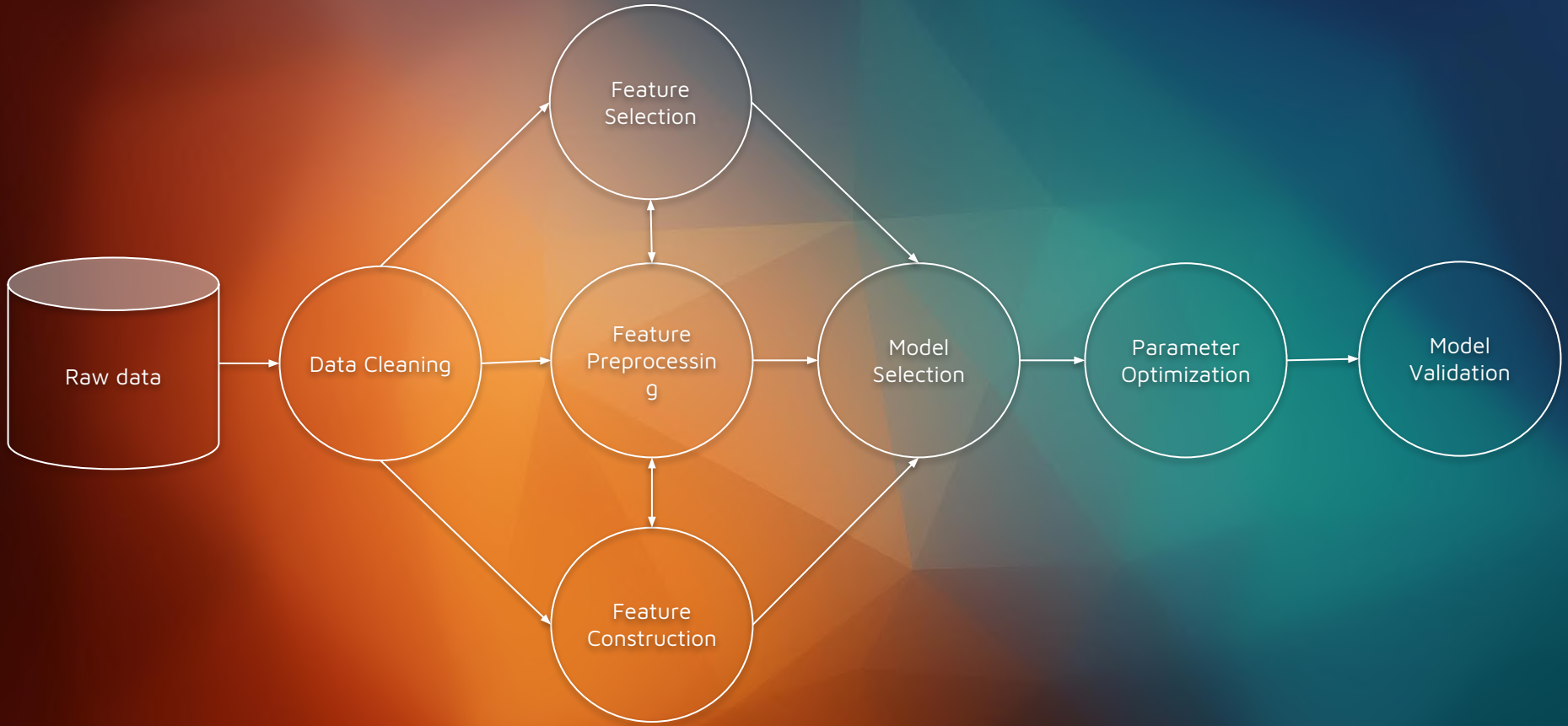
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=2, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```




Raw data
Data Cleaning
Feature Selection
Feature Preprocessing
Feature Construction
Model Selection
Parameter Optimization
Model Validation

Model validation




Feature Selection
Raw data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction

lazy

Oxford dictionary

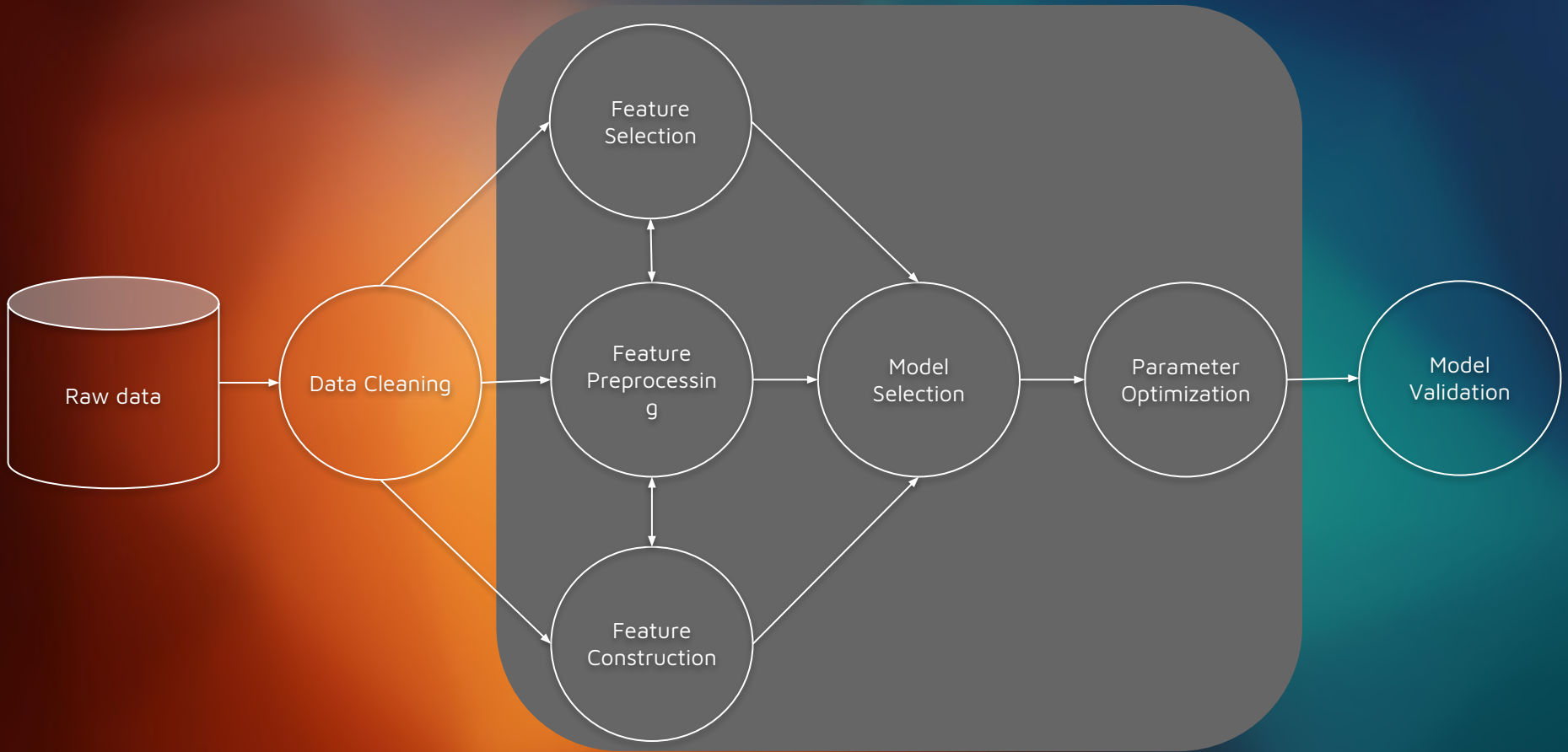
Unwilling to work or use energy

lazy

Diego dictionary

~~Oxford dictionary~~

Unwilling to work or use energy in repetitive tasks


Raw data
Data Cleaning
Feature Selection
Feature Preprocessing
Feature Construction
Model Selection
Parameter Optimization
Model Validation


naive_bayes.GaussianNB
tree.DecisionTreeClassifier
feature_selection.SelectFwe
ensemble.ExtraTreesRegressor
neighbors.KNeighborsRegressor
xgboost.XGBRegressor
ensemble.RandomForestRegressor
kernel_approximation.RBFSampler
preprocessing.MaxAbsScaler
linear_model.LogisticRegression
preprocessing.PolynomialFeatures
naive_bayes.MultinomialNB
feature_selection.SelectFromModel
preprocessing.RobustScaler
linear_model.ElasticNetCV
neighbors.KNeighborsClassifier
cluster.FeatureAgglomeration
preprocessing.Normalizer
feature_selection.VarianceThreshold
ensemble.GradientBoostingClassifier
decomposition.PCA
feature_selection.SelectPercentile
naive_bayes.BernoulliNB
svm.LinearSVR
ensemble.ExtraTreesClassifier
ensemble.AdaBoostRegressor
linear_model.LassoLarsCV
preprocessing.StandardScaler
tree.DecisionTreeRegressor
xgboost.XGBClassifier
tpot.builtins.ZeroCount
decomposition.FastICA
svm.LinearSVC
preprocessing.Binarizer
feature_selection.RFE
preprocessing.MinMaxScaler
linear_model.RidgeCV
kernel_approximation.Nystroem
ensemble.GradientBoostingRegressor
ensemble.RandomForestClassifier

Automated Machine Learning

TPOT

Auto-Sklearn

Auto-Weka

Machine-JS

DataRobot

auto_ml

Google Cloud AutoML

AutoKeras

H2O AutoML

...

TPOT is a Python tool that automatically creates and optimizes **machine learning pipelines** using **genetic programming**.

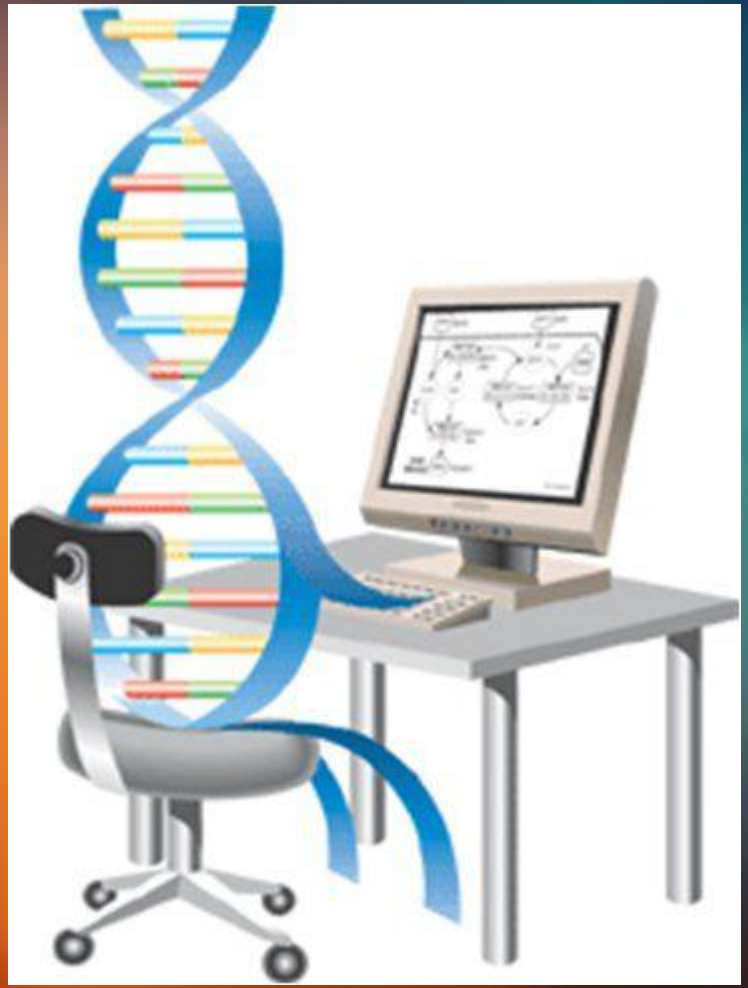
https://github.com/EpistasisLab/tpot

auto-sklearn

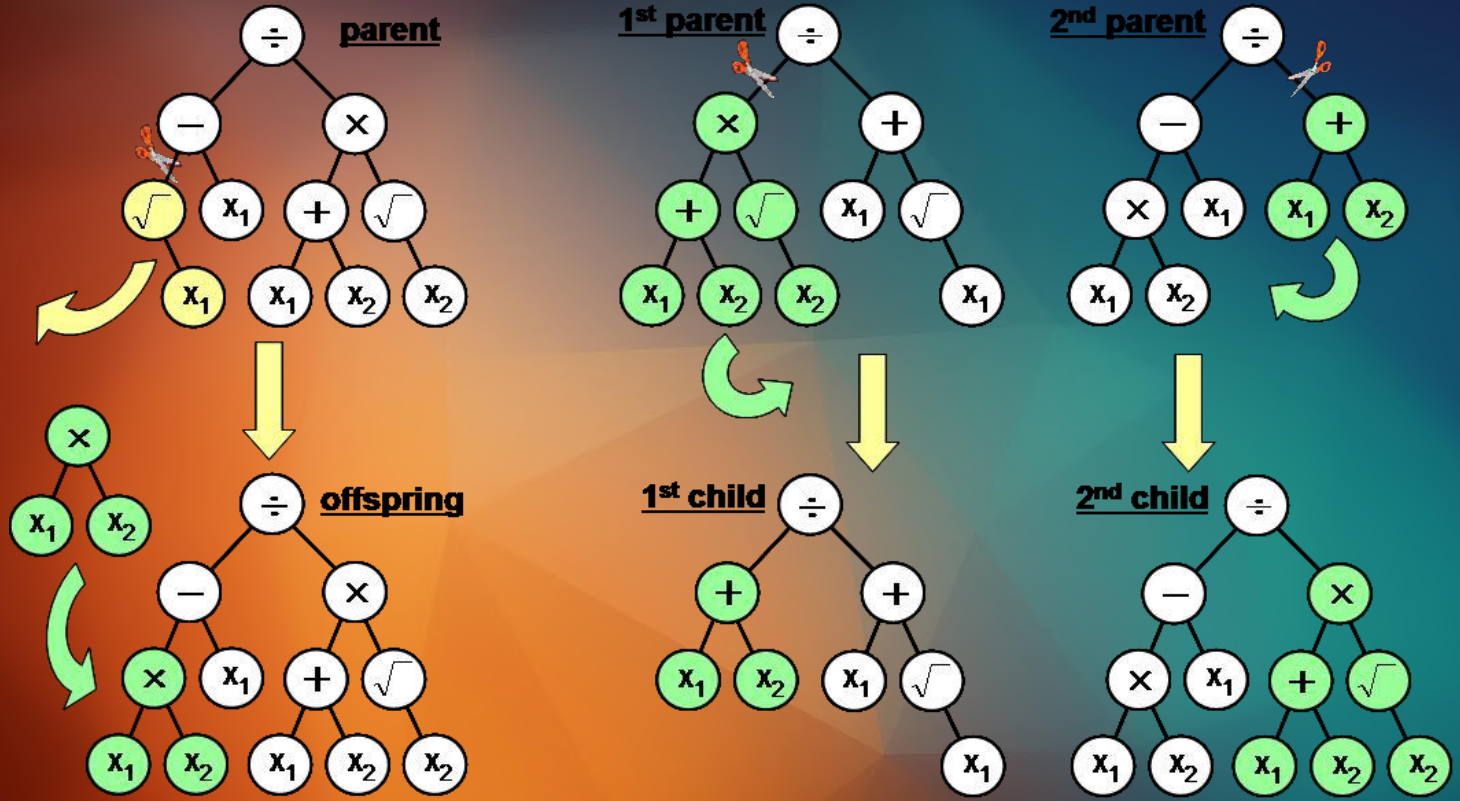
auto-sklearn frees a machine learning user from algorithm selection and hyperparameter tuning. It leverages recent advantages in **Bayesian optimization**, **meta-learning** and **ensemble construction**

https://github.com/automl/auto-sklearn

Genetic programming



Source: http://www.genetic-programming.org/gpbook4toc.html

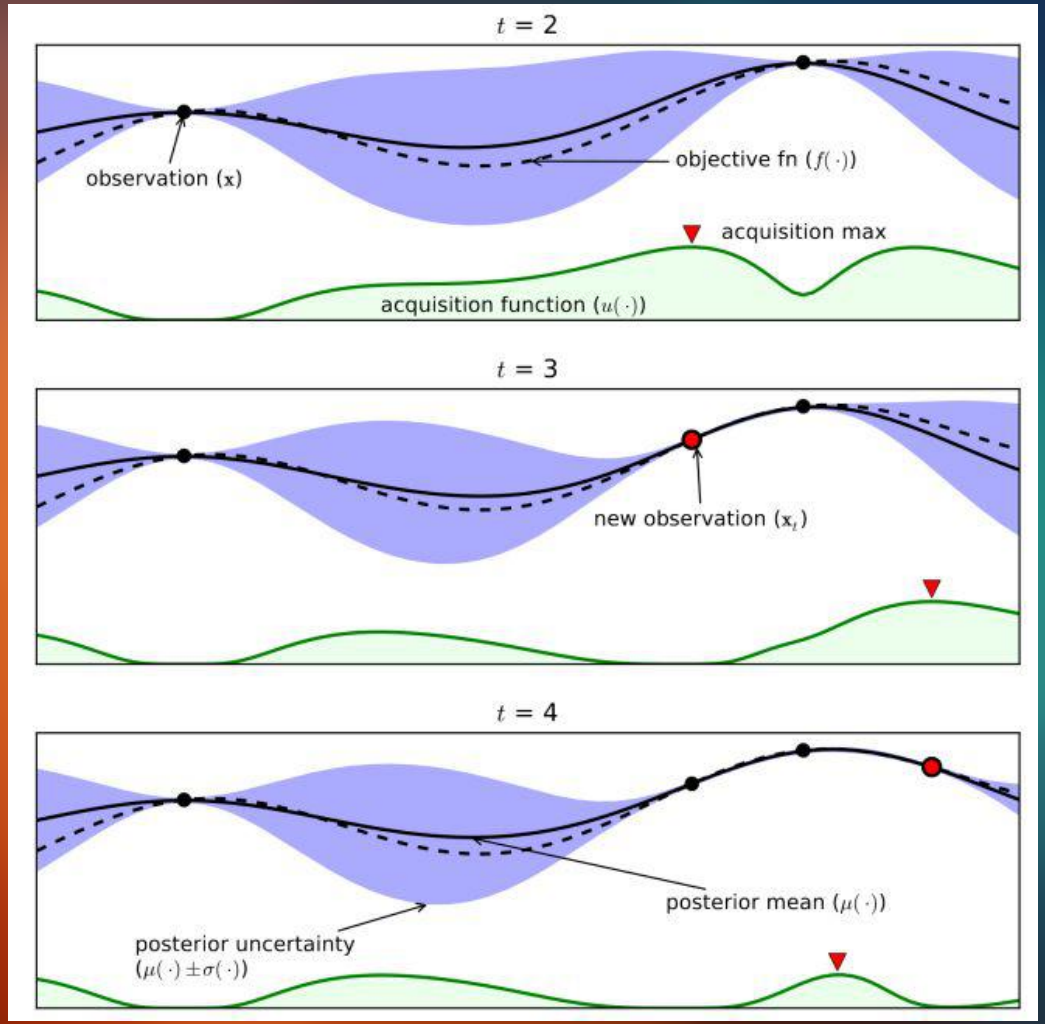


Mutation

Crossover

Source: http://w3.onera.fr/smac/?q=tracker

Bayesian optimization



source: https://advancedoptimizationatharvard.wordpress.com/2014/04/28/bayesian-optimization-part-ii/

TPOT is a Python tool that automatically creates and optimizes **machine learning pipelines** using **genetic programming**.

https://github.com/EpistasisLab/tpot

TPOT

```
from tpot import TPOTClassifier, TPOTRegressor


tpot = TPOTClassifier()
tpot.fit(X_train, y_train)
predictions = tpot.predict(X_test)


tpot = TPOTRegressor()
tpot.fit(X_train, y_train)
predictions = tpot.predict(X_test)
```

TPOT - configuration

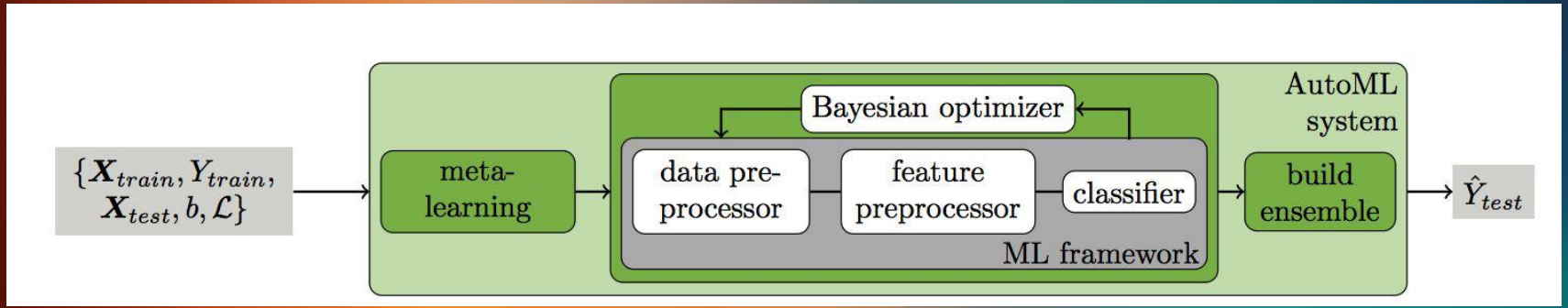
```
TPOTClassifier( config_dict  = {
   'sklearn.ensemble.RandomForestClassifier' : {
       'n_estimators' : [100],
       'criterion' : ["gini", "entropy"],
       'max_features' : np.arange( 0.05, 1.01, 0.05),
       'min_samples_split' : range(2, 21),
       'min_samples_leaf' :  range(1, 21),
       'bootstrap' : [True, False]
   },
   'sklearn.feature_selection.RFE' : {
       'step': np.arange( 0.05, 1.01, 0.05),
       'estimator' : {
           'sklearn.ensemble.ExtraTreesClassifier' : {
               'n_estimators' : [100],
               'criterion' : ['gini', 'entropy'],
               'max_features' : np.arange( 0.05, 1.01, 0.05)
              }
       }
   }
})
```

auto-sklearn

auto-sklearn frees a machine learning user from algorithm selection and hyperparameter tuning. It leverages recent advantages in **Bayesian optimization**, **meta-learning** and **ensemble construction**

https://github.com/automl/auto-sklearn

auto-sklearn



Source: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf

auto-sklearn

```
import autosklearn.classification
import autosklearn.regression

automl = autosklearn.classification.AutoSklearnClassifier()
automl.fit(X_train, y_train)
predictions = automl.predict(X_test)


automl = autosklearn.regression.AutoSklearnRegressor()
automl.fit(X_train, y_train)
predictions = automl.predict(X_test)
```

auto-sklearn custom config

include_estimators

exclude_estimators

include_preprocessors

exclude_preprocessors

Olive oil full dataset...

Test in Google Colab, clean dataset.

Tomorrow at 12:15

Hall 5

Olive oil full dataset...

Test in Google Colab, clean dataset.

TPOT: 55% Accuracy

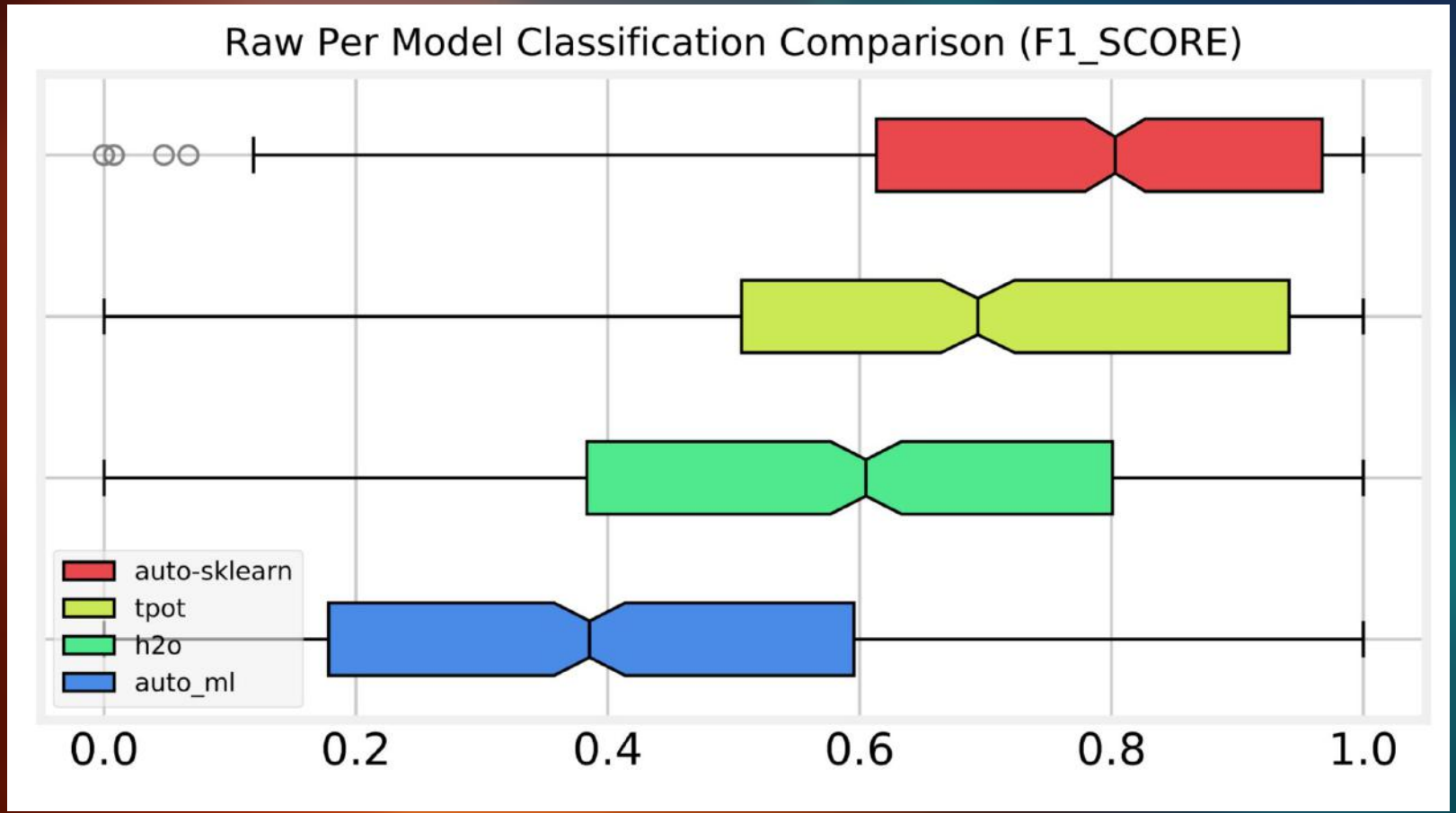
auto-sklearn: 56% Accuracy

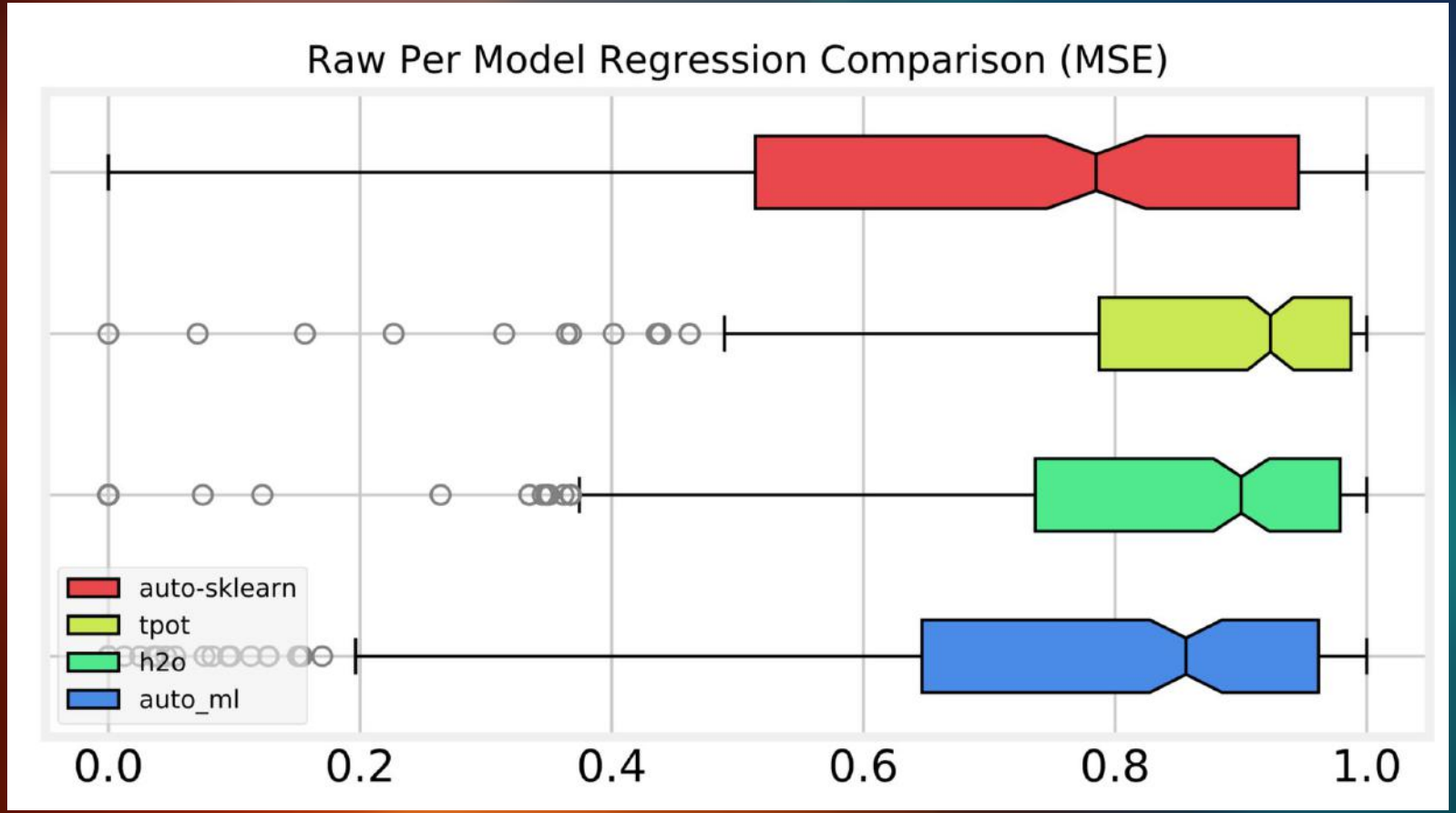
H2O automl: 51% Accuracy

Benchmarking Automatic Machine Learning Frameworks

Adithya Balaji, Alexander Allen

https://arxiv.org/abs/1808.06492


Raw Per Model Classification Comparison (F1_SCORE)
auto-sklearn
tpot
h2o
auto_ml
0.0
0.2
0.4
0.6
0.8
1.0


Raw Per Model Regression Comparison (MSE)
auto-sklearn
tpot
h2o
auto_ml
0.0
0.2
0.4
0.6
0.8
1.0

Automated Machine Learning—A Paradigm Shift That Accelerates Data Scientist Productivity @ Airbnb

https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8

Automated Machine Learning

- Exploratory analysis
- Selective discovering
- New ideas for your model
- Model optimization

Progress isn’t made by early risers. It’s made by lazy people trying to find easier ways to do something.

– Robert A. Heinlein

Thank you!

@jdiegoh

DiegoHueltes

Diego Hueltes

diego@hueltes.com

https://www.hueltes.com/automl